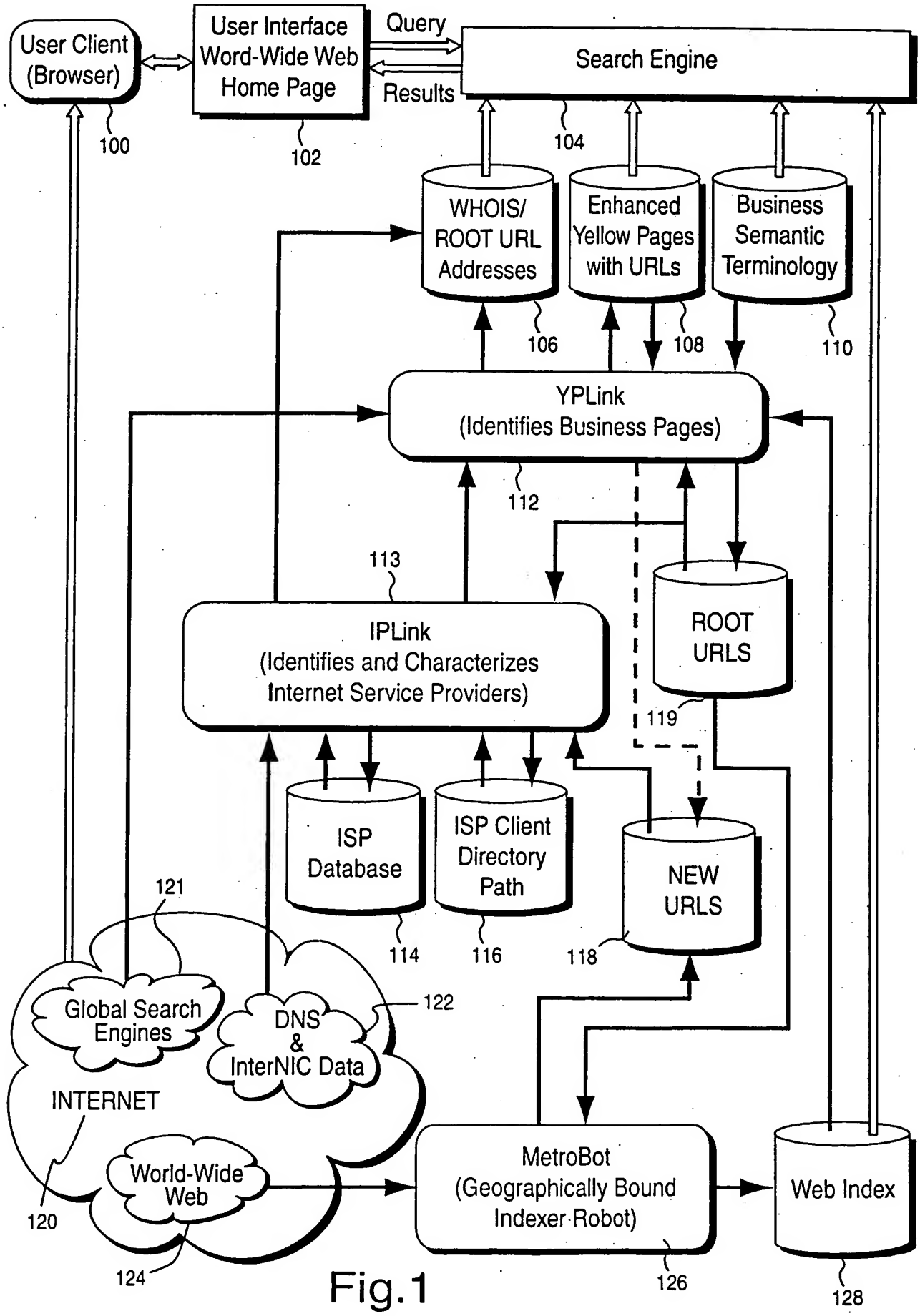
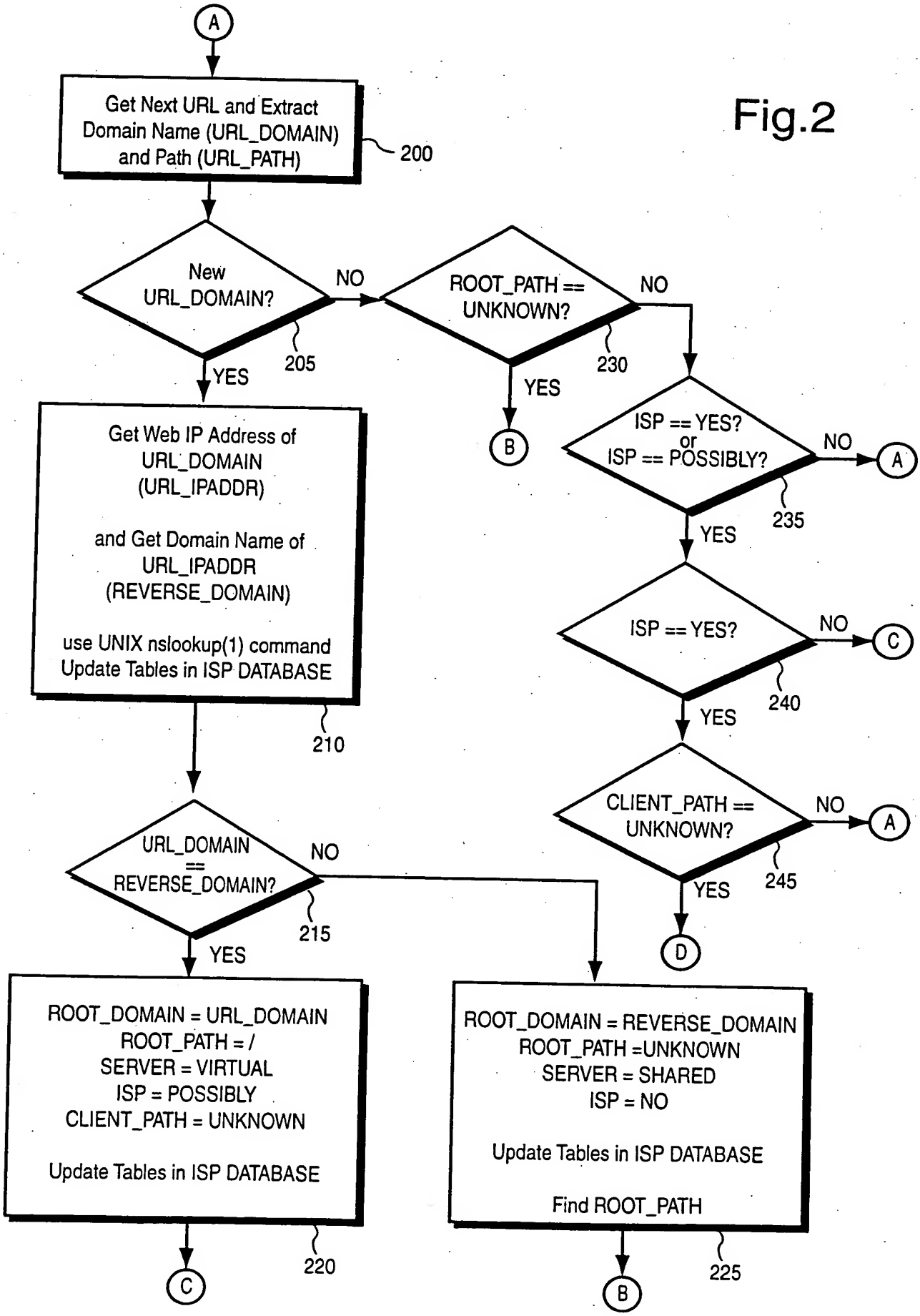

User Client (Browser)
User Interface Word-Wide Web Home Page
Query
Results
Search Engine
100
102
104
WHOIS/ ROOT URL Addresses
Enhanced Yellow Pages with URLs
Business Semantic Terminology
106
108
110
YPLink (Identifies Business Pages)
112
113
IPLink (Identifies and Characterizes Internet Service Providers)
ROOT URLS
119
ISP Database
ISP Client Directory Path
NEW URLS
114
116
118
121
122
Global Search Engines
DNS & InterNIC Data
INTERNET
World-Wide Web
120
124
MetroBot (Geographically Bound Indexer Robot)
126
Web Index
128

Fig.1

Fig.2


A
Get Next URL and Extract Domain Name (URL_DOMAIN) and Path (URL_PATH)
200
New URL_DOMAIN?
NO
205
YES
ROOT_PATH == UNKNOWN?
NO
230
YES
B
ISP == YES? or ISP == POSSIBLY?
NO
A
235
YES
ISP == YES?
NO
C
240
YES
CLIENT_PATH == UNKNOWN?
NO
A
245
YES
D
Get Web IP Address of URL_DOMAIN (URL_IPADDR)
and Get Domain Name of URL_IPADDR (REVERSE_DOMAIN)
use UNIX nslookup(1) command
Update Tables in ISP DATABASE
210
URL_DOMAIN == REVERSE_DOMAIN?
NO
215
YES
ROOT_DOMAIN = URL_DOMAIN
ROOT_PATH = /
SERVER = VIRTUAL
ISP = POSSIBLY
CLIENT_PATH = UNKNOWN
Update Tables in ISP DATABASE
220
C
ROOT_DOMAIN = REVERSE_DOMAIN
ROOT_PATH =UNKNOWN
SERVER = SHARED
ISP = NO
Update Tables in ISP DATABASE
Find ROOT_PATH
225
B

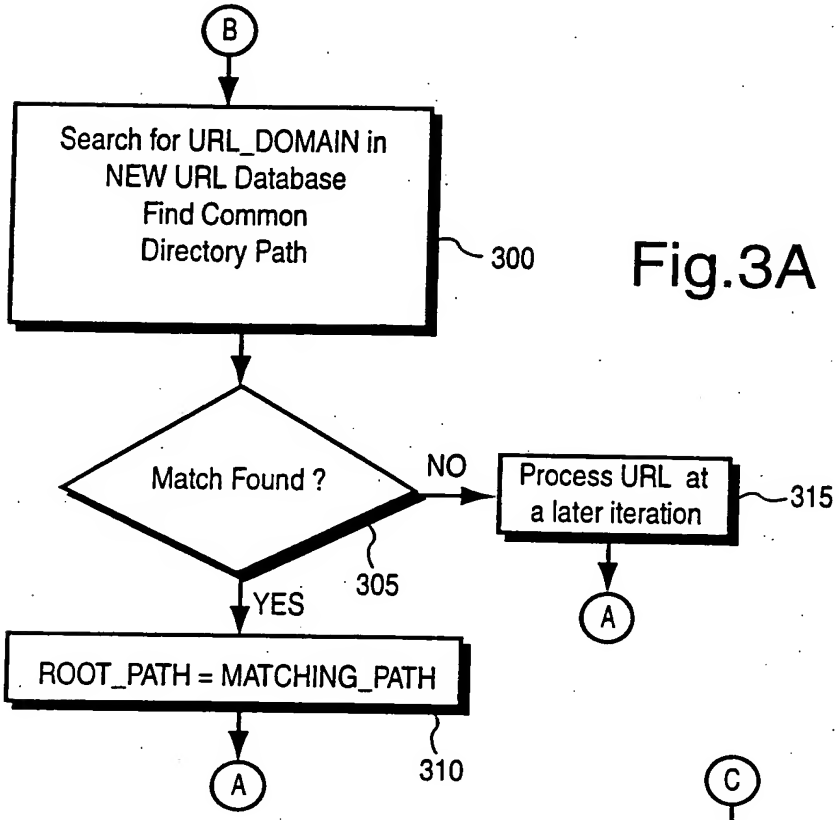


Fig.3A

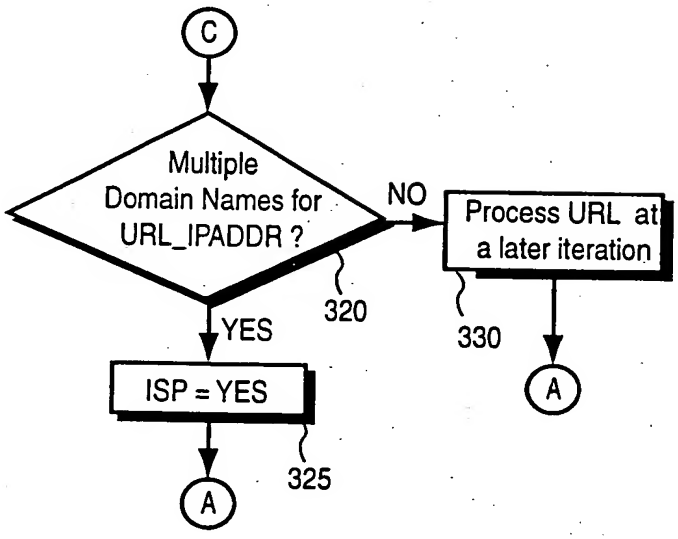


Fig.3B

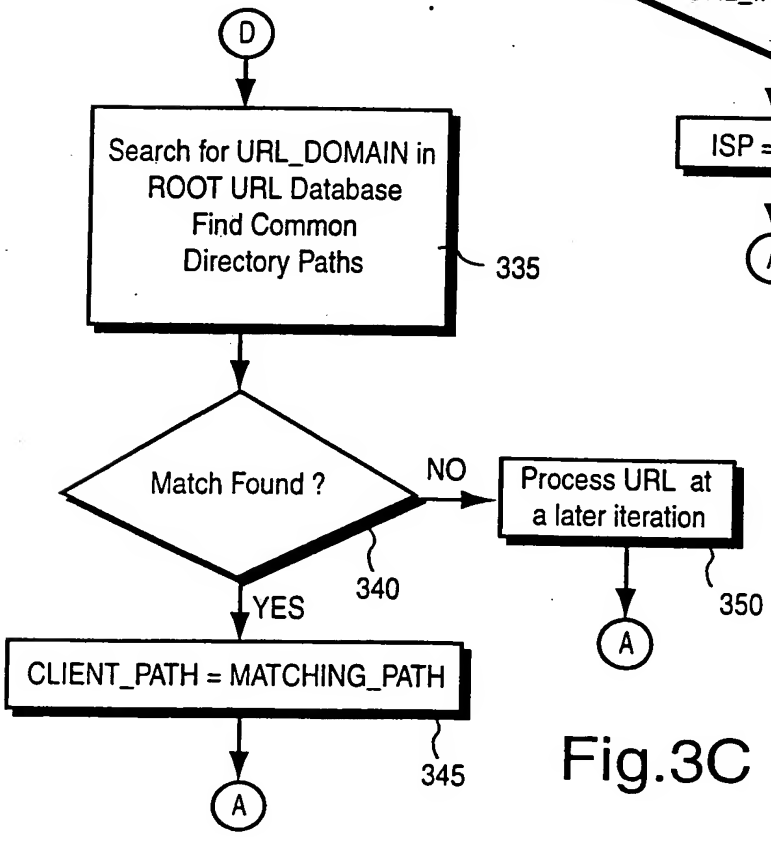


Fig.3C

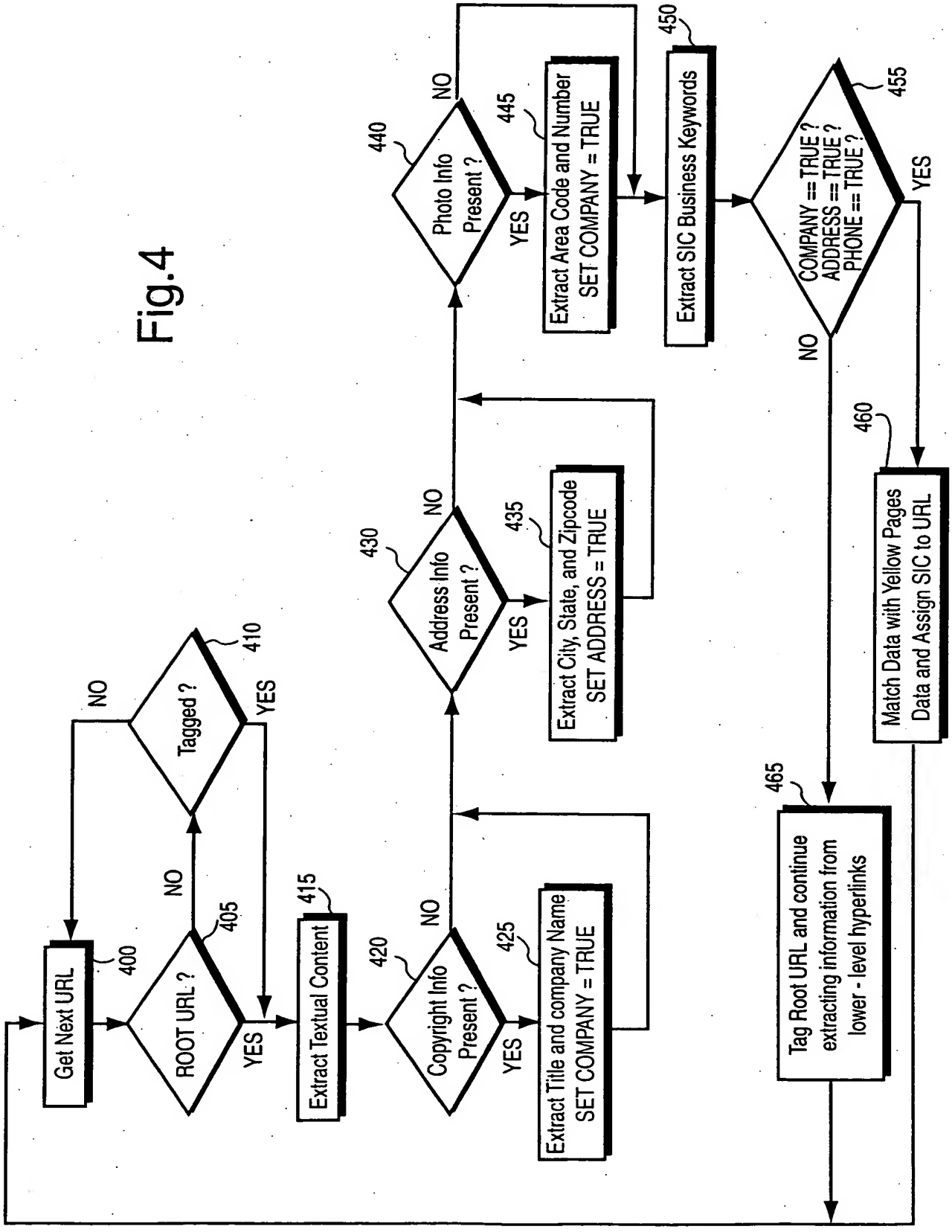

Get Next URL
400
ROOT URL ?
405
NO
YES
Tagged ?
410
NO
YES
Extract Textual Content
415
Copyright Info Present ?
420
NO
YES
Extract Title and company Name SET COMPANY = TRUE
425
Address Info Present ?
430
NO
YES
Extract City, State, and Zipcode SET ADDRESS = TRUE
435
Photo Info Present ?
440
NO
YES
Extract Area Code and Number SET COMPANY = TRUE
445
Extract SIC Business Keywords
450
COMPANY == TRUE ? ADDRESS == TRUE ? PHONE == TRUE ?
455
NO
YES
Match Data with Yellow Pages Data and Assign SIC to URL
460
Tag Root URL and continue extracting information from lower - level hyperlinks
465

Fig.4

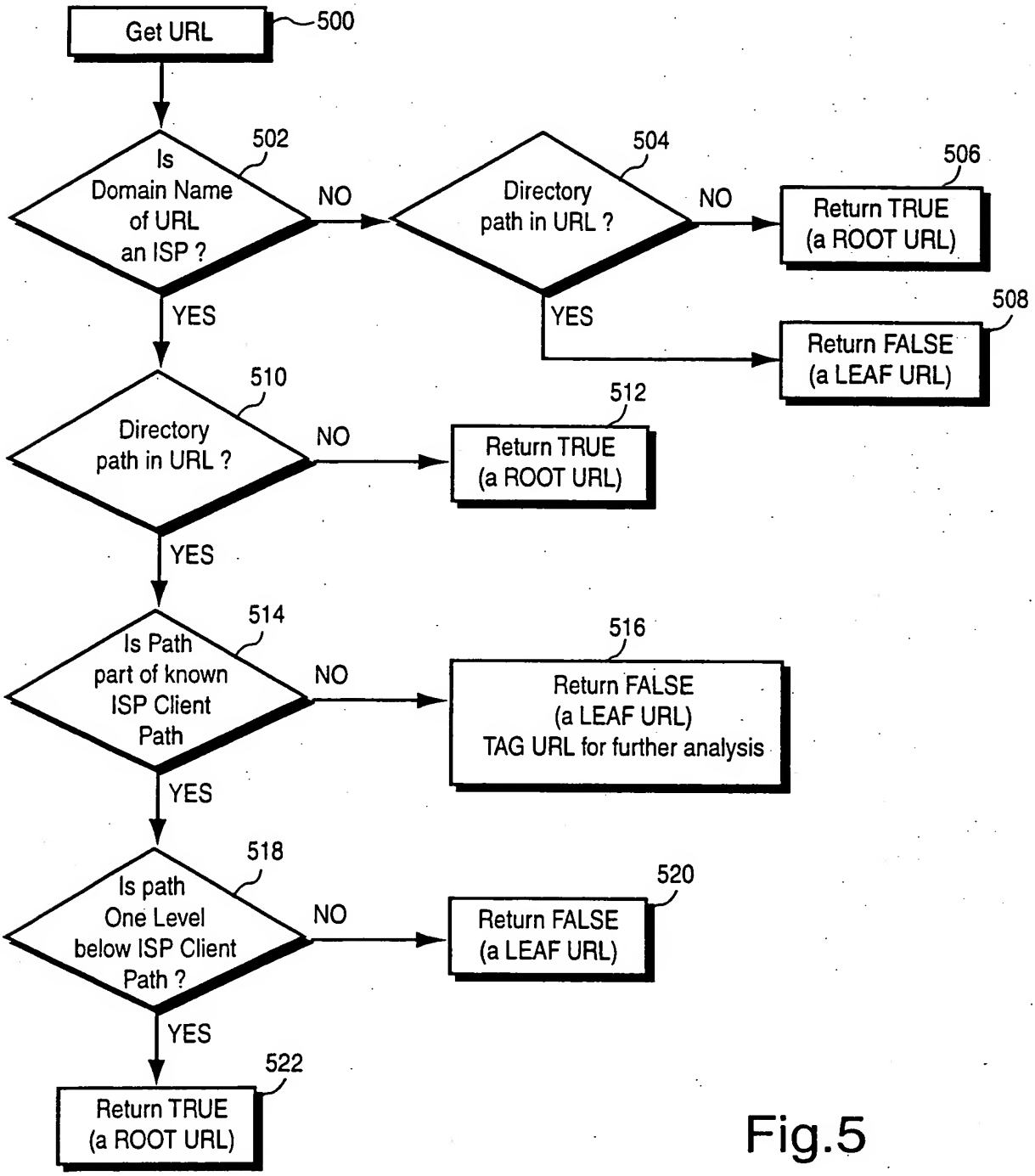

Get URL
500
Is Domain Name of URL an ISP ?
502
NO
Directory path in URL ?
504
NO
Return TRUE (a ROOT URL)
506
YES
YES
Return FALSE (a LEAF URL)
508
Directory path in URL ?
510
NO
Return TRUE (a ROOT URL)
512
YES
Is Path part of known ISP Client Path
514
NO
Return FALSE (a LEAF URL) TAG URL for further analysis
516
YES
Is path One Level below ISP Client Path ?
518
NO
Return FALSE (a LEAF URL)
520
YES
Return TRUE (a ROOT URL)
522

Fig.5

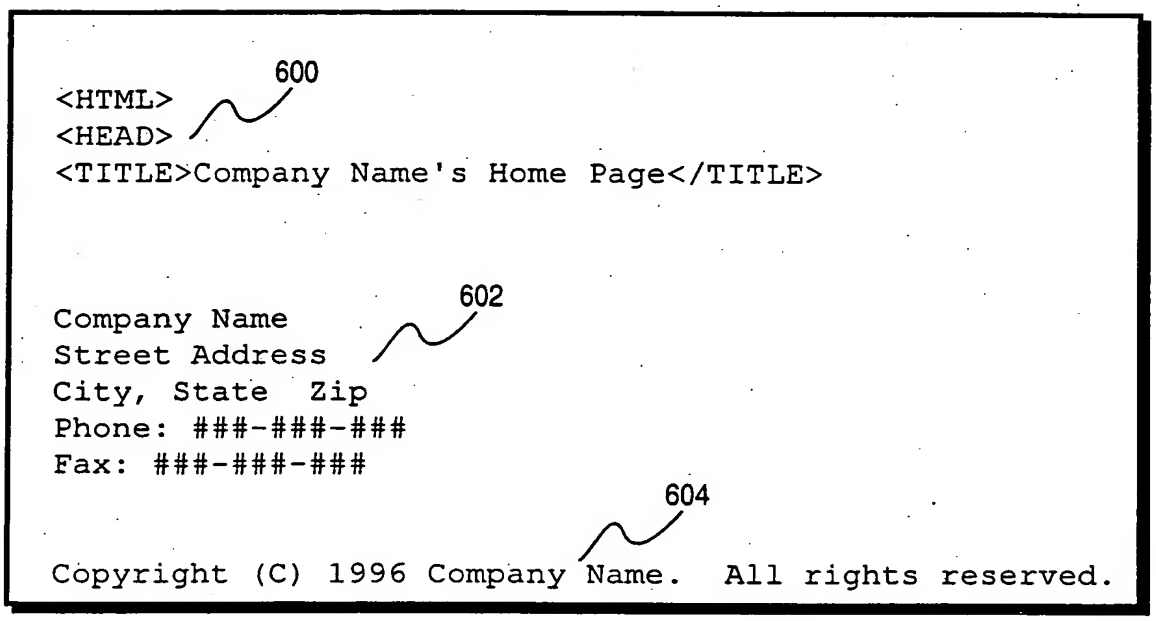

600
<HTML>
<HEAD>
<TITLE>Company Name's Home Page</TITLE>
602
Company Name
Street Address
City, State Zip
Phone: ###-###-###
Fax: ###-###-###
604
Copyright (C) 1996 Company Name. All rights reserved.

Fig.6